Electronic Thesauraus for Tamil

S.Baskaran
Sr.Programmer/Reader
Department of Computer Science
Tamil University
Thanjavur-613 004

sbaskarantnj@yahoo.com

S.Rajendran
Reader
Department of Linguistics
Tamil University
Thanjavur – 613 004

raj_ushush@yahoo.com

ABSTRACT

A thesaurus for Tamil is prepared by one the authors of this paper S.Rajendran. This thesaurus is used to create an electronic thesaurus for Tamil. The preparation of electronic thesaurus has some linguistics issues like hierarchical classification of lexical items, establishment of semantics domains, selection of lexical items, network of relations between lexical items, classification of lexical items under terminal domains etc. Computerization of thesaurus needs some procedures and methods for creating a special kind of databases and for accessing system. The thesaurus is prepared in roman scripts and provided facility to display the content in Tamil scripts also. The linguistics issues and computational procedures are dealt in this paper elaborately.

A thesaurus in its widest contemporary sense is a classification of words by concepts, topics, or subjects. The present electronic thesaurus is aimed to be built on the foundation offered by Natural Language Processing (NLP). There are some linguistic issues to be encountered and some computational methods to be adopted in preparing the electronic thesaurus.

Linguistic issues comprise of :

1. Hierarchical classification of lexical items based on at least four lexical relations: synonymy, hyponymy, compatibility, and incompatibility
2. Establishing semantic domains and sub domains based on distinguishing semantic or componential features

3. Selection of lexical items
4. Bringing out the network of relations existing between the lexical items which include co-occurrence and collocation
5. Arrangement of lexical items under terminal domains
6. Relating basic meaning with derivational meaning.

The preparation of electronic thesaurus, in general, needs the two procedures [1] Using computer corpus for preparing a thesaurus and [2] Creation of a database for building up a thesaurus. The software must provide the facilities for entering and editing the raw data, classifying the lexical items in semi-automatic way and user friendly interface for accessing the Electronic Thesaurus in simple manner.

A non-electronic thesaurus (paper thesaurus) for Tamil has been prepared by the present author (Rajendran, 2001). The preparation of electronic thesaurus for Tamil is an ongoing project.

1. Linguistic issues

1.1. Hierarchical Classification of lexical items

1.1.1. Synonymy, hyponymy, compatibility and incompatibility

A word acquires its referential meaning in being a member of a semantic domain by the common features it shares with other members in that domain, and having contrasting features which separate it from other members of the domain. It is the semantic relations among words, such as synonymy, hyponymy, compatibility and incompatibility, which help one to classify and organize words in terms of semantic domains in a hierarchical fashion. The following examples illustrates the semantic relations among lexical items.

SYNONYMY	Eg. *puttakam* 'book': *nduul* 'book'
HYPONYMY	Eg. *vilangku* 'animal': *pacu* 'cow', *erumai* 'buffalo'
COMPATIBILITY	Eg. *cellappiraaNi* 'pet': *ndaay* 'dog'

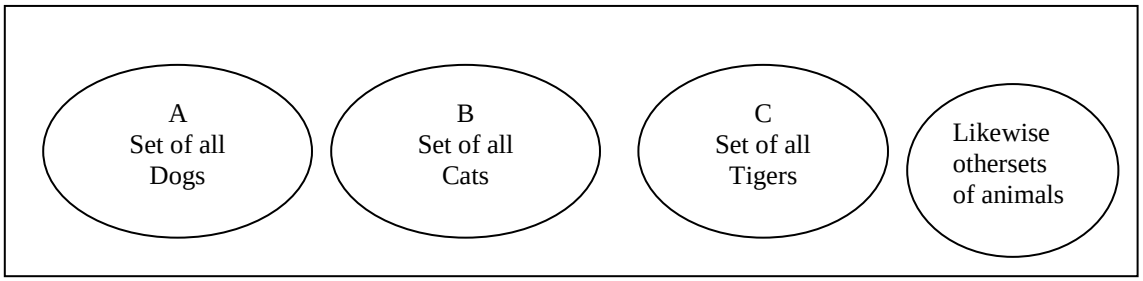
The relationship of compatibility can algebraically be expressed as follows.

Compatible: Let X = { Set of all dogs }
Let Y = { Set of all Pet Animals }
Compatible is a Function f between X and Y, $f : X \rightarrow Y$, in such a way that :
Then $\forall x$ in X there need not be an element y in Y such that y=f(x).
[All dogs need not be pet animals.]

For every y in Y there need be an x in X such that $x=f^{-1}(y)$.
[A pet animal need not be a dog.]

INCOMPATIBILITY Eg. *puunai* 'cat': *nfaay* 'dog'

Incompatible: Let X be a set of all disjoint subsets. X = { A,B,C .....}. Let Φ and Γ be any two elements of X , and Φ not = Γ and α be an arbitary element of Γ. Then Φ and Γ are incompatible if $\alpha \in \Gamma \Rightarrow \alpha \notin$ in Φ.



Let X be a set of all animals. Let A, B and C be a set of all dogs, cats, tigers, elephants respectively. Then $x \in A$ then x won't be either in B or in C. Here A,B and C are disjoint subsets. They are said to be incompatible with each other.

1.1.2. Binary opposition

Antonymy, which is often considered as opposite to synonymy, relies on the lexical relation of incompatibility. Lyons (1977:270-290) reserves the term antonymy for a narrow sense, and describes several different kinds of lexical 'opposites' likes antonymy, complementarity, Private opposition and equipollent opposition. The following examples illustartes them.

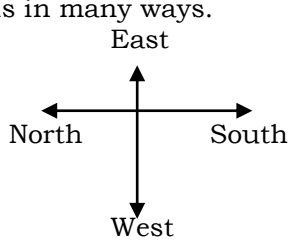
ANTONYMY Eg. *peritu* 'big':*ciRitu* 'small'
COMPLEMENTARITY Eg. *pakal* 'day': *iravu* 'Night'
PRIVATIVE OPPOSITION Eg. *aakfRiNai* 'irrational':*uyartiNai* 'rational'
EQUIPOLLENT OPPOSITION E.g. *aaN* 'male':*peN* 'female'

A further opposition to be distinguished from antonymy and complementarity is converseness, exemplified by such pairs as *kaNavan* 'husband': *manaivi* 'wife'. Converses typically include reciprocal social roles, kinship relations, temporal relations, spatial relations and complex relations.
CONVERSENESS
Reciprocal Social roles Eg. *maruttuvar* 'doctor':*ndooyaaLi* 'patient'
Kinship Relations Eg. *Appaa* 'father': *makan* 'son', *ammaa* 'mother':*makaL* 'daughter'

Temporal Relations Eg. *munnar* 'before':*pinnar* 'after'
Spatial Relations Eg. *meelee* 'above':*kiizhee* 'below'
Complex Relations Eg. *vaangku* 'buy':*vil* 'sell'

In addition, there can be defined three further oppositions, those of directional opposites orthogonal opposites and antipodal opposites. The vocabulary of kinship and color can also manifest the principles of orthogonal and antipodal oppositions in many ways.

DIRECTIONAL OPPOSITION Eg. *ceer* 'arrive':*puRappaTu* 'depart'
vaa 'come':*poo* 'go'



ORTHOGONAL OPPOSITION or PERPENDICULAR OPPOSITION
Eg. *vaTakku* 'north':*kizhakku* 'east' and *meeRku* 'west'
kizhakku 'east':*teRku* 'south' and *vaTakku* 'north'
ANTIPODAL OPPOSITION or DIAMETRICAL OPPOSITION
Eg. *vaTakku* 'north':*teRku* 'south', *kizhakku* 'east:*meeRku* 'west'

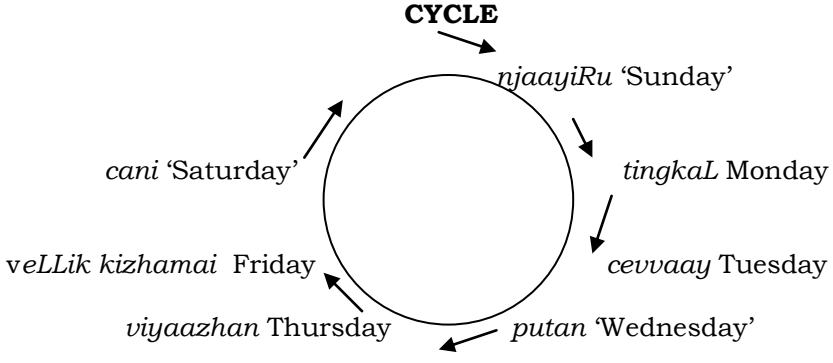
1.1.3. Multi-member opposition

There are different types of multi-member sets in a language whose lexical relations can be described as incompatible denoting non-binary contrasts as opposed to binary contrasts. Various kinds of ordering are found in multi-member sets of incompatibles, and such sets may be serially or cyclically ordered. The ordering in scales in terms of incompatibility is typically less strict than it in ranks. The examples given below will stand to explain the difference between them.

SERIAL

ndaaTu 'country'
|
maanilam 'state'
|
maavaTTam 'district'
|
vaTTam 'thaluk'
|
kiraamam 'village'

CYCLE



SCALE

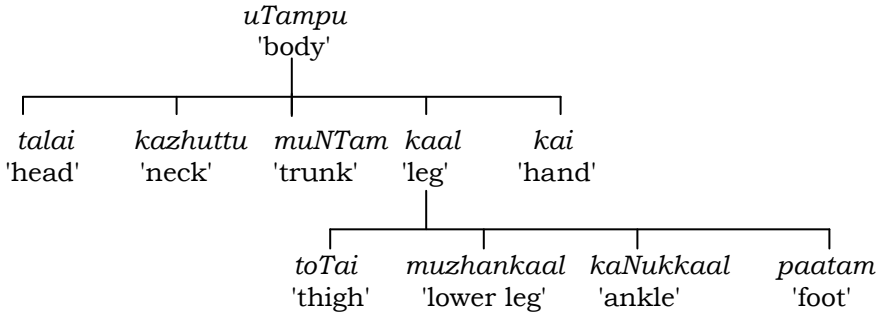
mikandanRu 'excellent'
|
ndanRu 'good'
|
tirupti 'satisfactory'
|
moocam 'bad'
|
mika moocam 'very bad

RANK

onRu 'one'
|
iraNTu 'two'
|
muunRu 'three'
|
ndaanku 'four'
|
aindtu 'five'

1.1.4. Meronymy

Meronymy (part-whole relation) also plays an important role in the hierarchical arrangement of lexical items. The division of the human body into parts can serve as a prototype for all part-whole hierarchies:



The fact that the lexical items of a language fall into a number of distinct parts of speech would itself preclude the hierarchical ordering of the vocabulary of a language in terms of hyponymy and hypernymy. But it is difficult to assign members of a parts of speech under a super ordinate term or even a set of items under a super ordinate term.

1.1.5. Nida's hierarchical classification

Nida (1975a) who was concerned with the preparation of a thesauric dictionary for Greek gives the following as the tentative hierarchical classification of the lexical items (Nida:178-186).

I. Entities
  A. Inanimate
    1. Natural
      a. Geographical
      b. Natural substances

c. Flora and plant products
2. Manufactured or constructed entities
a. Artifacts (non constructions)
b. Processed substances: foods, medicines, and perfumes
c. Constructions
B. Animate entities
1. Animals, birds, insects
2. Human beings
3. Supernatural power or beings

II. Events

A. Physical, B. Physiological, C. Sensory, D. Emotive, E. Intellection, G. Communication, G. Association, H. Control, I. Movement, J. Impact, K. Transfer, L. Complex activities, involving a series of movements or actions

III. Abstracts

A. Time, B. Distance, C. Volume, D. Velocity, E. Temperature, F. Color, G. Number, H. Status, I. Religious character, J. Attractiveness, K. Age, L. Truth-falsehood, M. Good-bad, N. Capacity, O. State of health, etc.

IV. Relational A. Spatial, B. Temporal, C. Deictic , D. Logical, etc.

This classification is based on referential meanings and it is not possible to obtain one to one correspondence between the semantic domain of classes and the grammatical classes.

1.1.6. Scientific taxonomy vs. folk taxonomy

There is often a clash between scientific taxonomy and folk taxonomy, particularly in such areas as those of flora and fauna. As Tamil is aiming at equipping itself for scientific use, preference may be felt for a scientific classification over that enshrined in folk taxonomy which reflect culture, in a thesaurus of modern Tamil. "Modern scientific taxonomy excludes *iraal* 'shrimp' from its traditional place in the class of fish, and classifies *vauvaal* 'bat' not as a bird but as an animal. We would wish to place *puumi* 'earth' and *candtiran* 'moon' separately from *ndaTcattirankaL* 'stars' and *cuuriyan* 'sun' along with 'stars', and to eliminate *iraaku* 'a mythical serpent believed to consume the sun and the moon and cause eclipses' and *keetu* 'a mythical celestial dragon believed to cause eclipses' from the class of astronomical phenomena. Preference for astronomy over astrology reflects a contemporary scientific approach to thesaurus construction." (Rajendran, 1995:84-85)

1.1.7. Superordinate term

Finding suitable superordinate term or heading for a group of items often poses problem. The availability of a super ordinate term which can cover a semantic domain facilitates the assigning of membership to the domain.

1.2. Establishing semantic domains and sub domains

The hierarchical structuring of lexical items and establishing boundary between domains can sometimes be difficult. One cannot always except a sharply defined and dichotomously distinguished neat classification. Also, though it may be easy to decide about the inclusion of items which form the core of a domain, there may be many border-line cases where the decision of inclusion becomes difficult.

1.3. Selection of lexical items

Selection and inclusion of disputable lexical items envisages stylistic alternants, derived forms, idioms and phrases, compounds, proper names (which includes names and epithets of gods and goddess), names of plants and animals and obsolete usages. Rajendran (1995:88-92) has discussed in details about the problems involved in the selection and inclusion of the above mentioned items in a thesaurus. Generally speaking it is the size of the thesaurus which will decide the nature of the items to be incorporated in a thesaurus, though there is no dispute about the inclusion of lexical items one may find in a general dictionary meant for language users.

STYLISTIC ALTERNANTS Eg. king: *raajaa, raacaa, iraacaa*
DERIVED FORMS Eg. student: *maaNavar; maaNavan; maaNavi*
IDIOMS & PHRASES Eg. covertly: *ilai maRai kaay maRaivaaka, terintum teriyaamalum*
COMPOUNDS Eg. *maavilai* 'mango leaf'; *maampuu* 'mango flower'; *maampazham* 'mango fruit';... |
PROPER NAMES & EPITHETS Eg. Names of Lord Siva: *ndandti, pommaan, aalamarndtoon, ndampan, iRai, cooti, cemmal, andtivaNNan, taaNu,.. .*

1.4. Bringing out the net work of lexical relations

An electronic thesaurus should bring out the semantic relations existing between the lexical items. Though it may be difficult to show all these relations in a paper thesaurus, it is possible to bring out this relations in an electronic thesaurus for the benefit of its users in a friendly fashion.

1.5. Arrangement of lexical items

Though arrangement of lexical items under a terminal domain is not a serious issue in an electronic thesaurus, the decision of what to give first among a set of synonyms is a problem to be encountered in the preparation of the thesaurus. Criteria based on core meaning vs. peripheral meaning, higher style vs. lower style and more frequently used vs. less frequently used may help in deciding the priority in the arrangement of synonyms.
sky: *vaanam, aakaayam, aakaacam* (s.v.), *vaan* (h.v.), viN (h.v.); *meel vaanam; kiizh vaanam* [s.v. = spoken variety, h.v. = high variety]

1.6. Relating basic meaning with derived meaning

Nida gives derivational meanings under their basic meaning, i.e. derived items under their bases. Nida's reason for placing derived meanings under their bases is to assign componential features to the derivatives from the bases, thus avoiding duplication. But this can be done without duplication by cross reference. Listing the derived items under the class to which they belong by virtue of their derived meanings and furnishing with cross reference for the bases from which they are derived seem to be a better solution. This can be easily accomplished in an electronic thesaurus.

azhaku 'beauty'
azhakan 'handsome person'; *azhaki* 'beautiful woman'
azhakaana 'beautiful'
azhakaaka 'well'

2. Computational issues

2.1. Computer Corpus

A computer corpus, as we understand, is a large body of naturally occurring computer-readable texts or text extracts used for research, and especially for the development of natural language processing. Any work related to lexicon presumes a computer corpus these days. Using a computer corpus for lexical work makes things easy. Also the corpus helps in making any decision with maximum perfectness. A corpus annotated for grammatical categories and semantic information will be a useful tool in thesaurus making. Automatic tagging for grammatical information is feasible, but automatic tagging for semantic information needs a knowledge base, the building up of which is a difficult task.

2.2. Creation of a database

Corpus can be made use of for the creation of database for making a thesaurus. A tagged corpus can help us to prepare a thesaurus more effectively. Dictionaries can be used as a secondary source. The paper thesaurus prepared by Rajendran (2001) can be used to create the database for the preparation of electronic thesaurus. The hierarchical classification of vocabulary of Tamil is available in the above mentioned paper thesaurus.

The concepts of Relational Data Base Management System [RDBMS] is employed for creating and maintaining of the database of electronic thesaurus. For searching, the concept of Backtracking Method is adopted in the context when backtracking is a mandatory.

2.3. Providing user friendly interface

A user friendly interface helps the user to get the needed information with great ease. This includes the following issues: [1] Transfer of classificatory and network scheme into commands for electronic type setting, [2] Transfer of classificatory and network scheme into information retrieval system[3]On-line retrieval: combining the classificatory and network scheme with code searching.

REFERENCES

[Cruse, 1986] Cruse, D.A. 1986. Lexical Semantics. Cambridge: Cambridge University Press.

[Guckler, 1983] Guckler, G. 1983. Appendix: B: A Computer-based Monolingual Dictionary: A Case Study. In R.R.K. Hartmann (ed.) 1983. Lexicography: Principles and Practice. London: Academic Press Inc.

[Jones, 1986] Jones, K.S. 1986. Synonymy and Semantic Classification. Edinburgh: Edinburgh University Press.

[Martin et al., 1983] Martin, W.J.R., B.P.F.Al and P.J.G.Van Sterkenburg. 1983. On Processing of A text Corpus. In R.R.K. Hartmann (ed.). 1983. Lexicography: Principles and Practice. London: Academic Press Inc.

[Nida, 1975] Nida, E.A. 1975a. Compositional Analysis of Meaning: An Introduction to Semantic Structure. The Hague: Mouton.

---------. 1975.b. Exploring Semantic Structure. The Hague: Mouton

[Rajendran, 1978] Rajendran, S. 1978. Syntax and Semantics of Tamil Verbs. Ph.D. Thesis. Poona: University of Poona.

[Rajendran, 1982] Rajendran, S.1982. Semantics of Tamil Vocabulary. Post-doctoral Research Report. Poona: Deccan College.

--------. 2001. *taRkaalattamizhc coRkaLanciyam* [Thesaurus for Modern Tamil]. Thanjavur: Tamil University.

--------. 1995. "Towards a Compilation of a Thesaurus for Modern Tamil". South Asian Language Review. 5.1:62-99

--------. 1996. "The Feasibility of Preparing a Thesaurus using Corpus". Workshop on Indian Language Corpus and its Applications (28, 29 Oct. 1996). Central Institute of Indain Languages, Mysore.